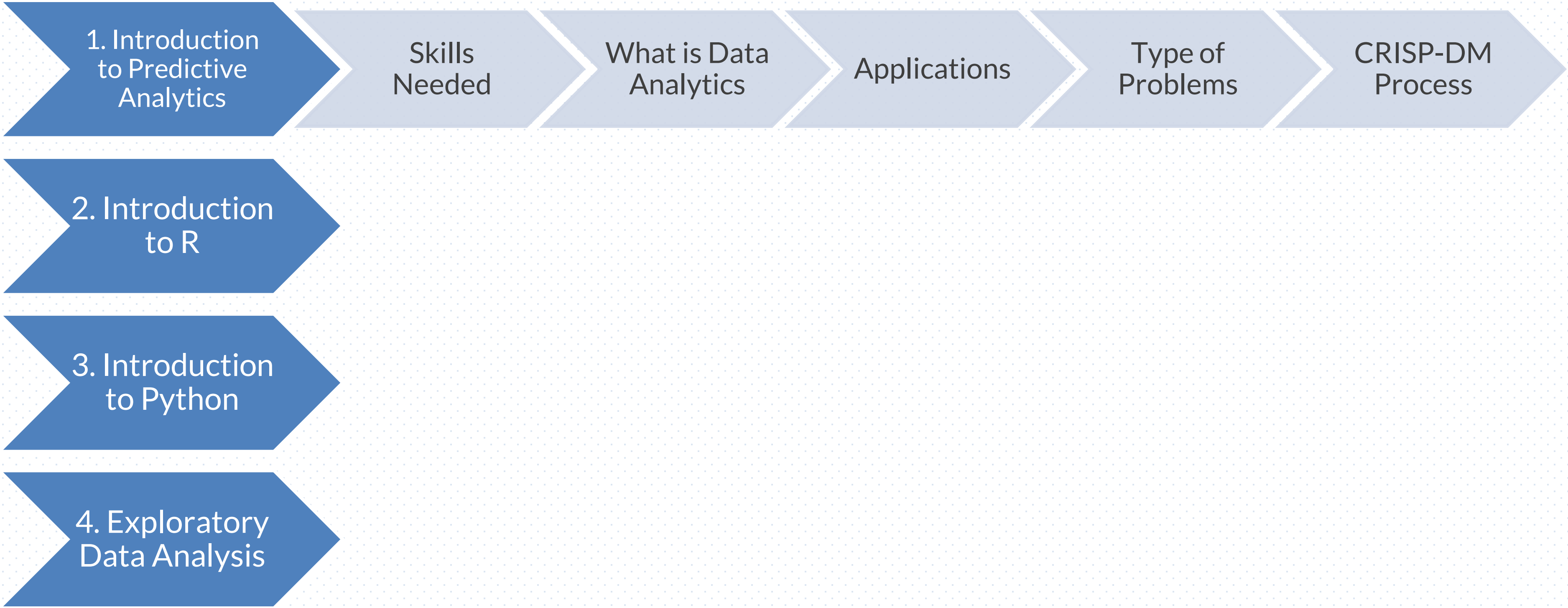
Predictive Analytics Using R and Python

- MUQUAYYAR AHMED
DATA SCIENTIST

We learnt!!



1. Introduction to Predictive Analytics

Skills Needed

What is Data Analytics

Applications

Type of Problems

CRISP-DM Process

2. Introduction to R

3. Introduction to Python

4. Exploratory Data Analysis

Today’s Agenda

- Missing Value Analysis

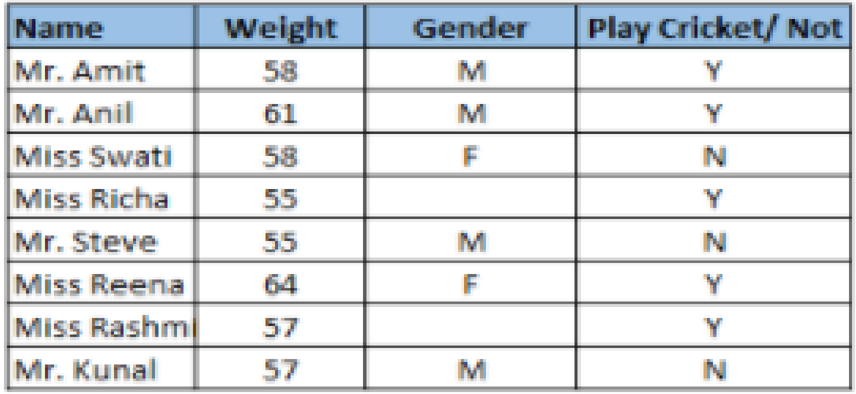
Missing value Analysis

Why missing values

- Human Error
- Refuse to answer while surveying
- Optional box in questionnaire

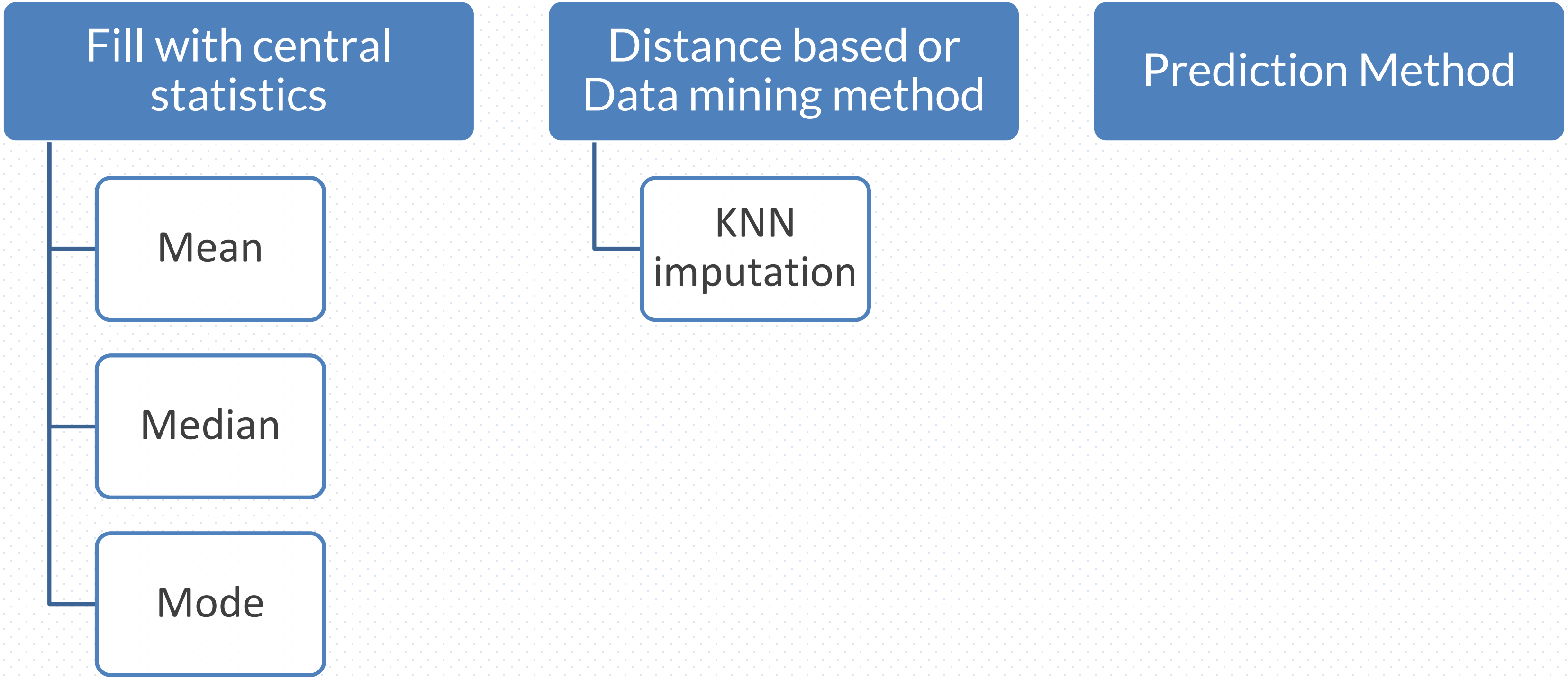
Ignore or impute missing value???

- Understand why each value is missing
- Plot bar graph
- Delete observations or variables where you do not intend to impute a value
  - ○ Drop variable
  - ○ Drop observation
  - ○ Consider the variables to impute whose missing values is less than 30%



Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Impute missing values



Fill with central statistics

- Mean
- Median
- Mode

Distance based or Data mining method

- KNN imputation

Prediction Method

KNN-Imputation

- Find the nearest neighbor based on existing attributes
- Use Euclidean or Manhattan distance
- Euclidean distance

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

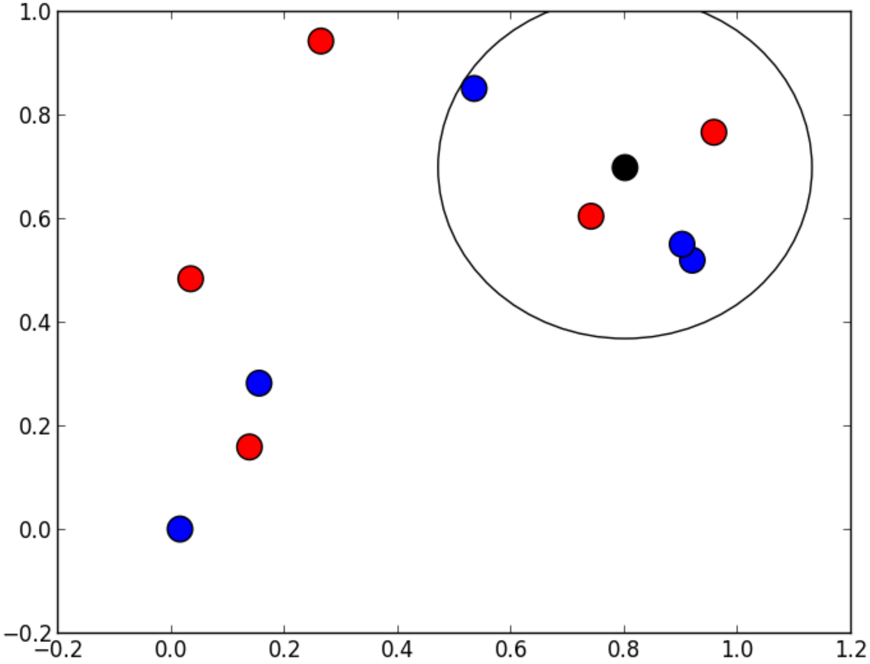
	a_1	a_2	a_3	a_4
R_1	( )		( )	( )
R_2	( )	NA	( )	( )
R_3				
R_4				

$$(x_2 - x_1)^2$$

Contd..

- **Take an average of only the nearest neighbors**
  - –Mean for numeric
  - –Mode for categorical

- **K should be odd for categorical variable**



Framework

- Create a small subset of data with complete observations
- Delete some values manually
- Use multiple methods to fill
- See where they are failing
- Choose the best method